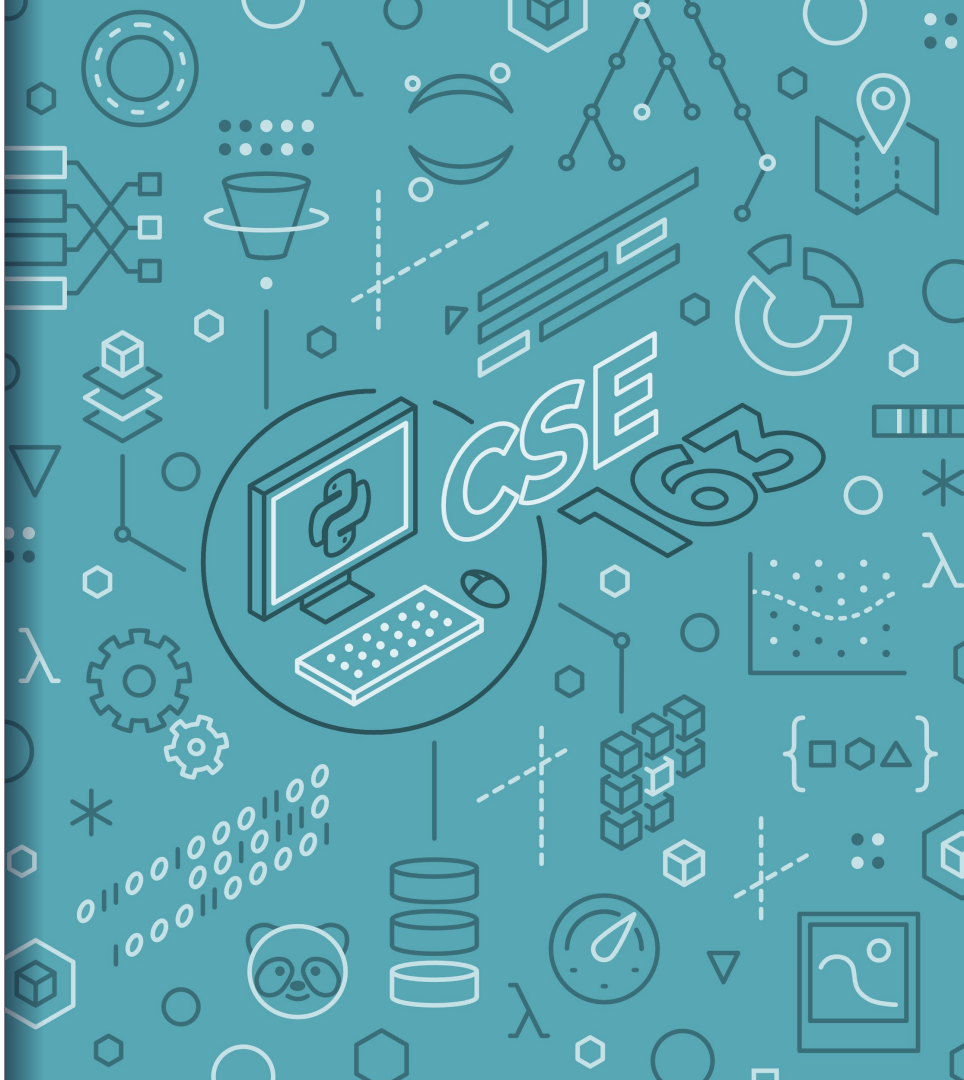
CSE 163

Web Scraping

Hunter Schafer



Logistics

- Exam 2 scores will be published by Monday morning

- Project Part 2 is due Friday night
  - Select partners correctly on Gradescope
  - Make sure to turn in both report, code, and data (if possible, instructions to get data if not)
  - Indicate Live Presentation or Video for Project Part 3
- Project Part 3
  - Resources due next Tuesday at **noon.**
    - **No late submissions.**
  - Live presentations during the final exam slot Tuesday 2:30-4:20
  - Everyone must be there (regardless if presenting or not)

Internet

Allows computers to talk to each other to communicate data

One way to use the internet is to use a browser (e.g. Google Chrome) to browse web-pages.

Everything is determined by a URL that identifies

- Which computers to talk to
- Which resource you want that computer to give you

https://courses.cs.washington.edu/courses/cse163/19sp/files/hw7/images/puppy.png

Resources

In most cases, a URL specifies either

- A file that we want to access
  - Example: https://www.google.com/imghp?hl=en&tab=wi&authuser=0&ogbl
- A special command to the particular service to return some meaningful information. This is called an **Application programming interface (API).**
  - Example: http://api.open-notify.org/iss-now.json
  - There are many types of APIs that return different things depending on what you ask, but it is a useful way for one computer to get data from another
  -

API Demo

- requests module
- GET request
  - Response codes
  - Response data
- JSON data

For this example, we used the Open-Notify API

- Documentation: http://open-notify.org/Open-Notify-API/

Colab

HTML

A website is actually just text that's in a very special format: **HTML**

Basic idea:

- Use "tags" around parts of the text to indicate meaning
- Example tags
  - Paragraphs
  - Images
  - Links
  - Etc.

HTML Tags

Generally, we use an open tag and a close tag to say indicate everything inside belongs to that type of data:

- Tags can have attributes (src, href)
- Some tags are “self-closing” like img

```
<p>This is a paragraph.</p>

Here is an image: <img src=”dog.jpg”>

<p>
    This is another paragraph but it has a
    <a href=”http://www.puppies.com”>link</a>!
</p>
```

Wikipedia > View Source

Brain Break

Web Scraping

Looking at the HTML of a page in order to extract meaningful information from it.

You could imagine trying to parse this would be a nightmare!

There is a library to do that for us though! We will use Beautiful Soup (bs4) to do all the nast HTML parsing for us.

For the example today, we will be looking at the weather in Seattle.

Finding Elements

Looking for all paragraphs would be overly tedious since there might be many paragraphs and only some are useful.

HTML lets you use “user defined” names for tags.

- IDs (a tag can have at most one ID that must be unique)
- Classes (a tag can have multiple classes and classes don’t have to be unique)

```
<p id="introduction" class="fancy">
    This is a paragraph.
</p>

<p class="fancy warning-text">
    This is another paragraph!
</p>
```

Web Scraping

Demo

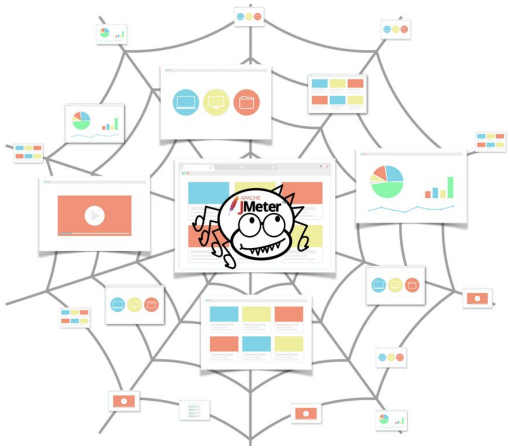
Goal: Use Beautiful Soup to extract information about the weather forecast

Web Crawler

Our demo was limited because we were just looking at one page

The power of the internet is that pages link to other pages!

You can take this another step further to “crawl” the web by visiting sites that are linked to by your current site.



Web Scraping

Good

- Can work for almost any website! All you have to do is know what the structure of the web-page is and you can extract anything!

Bad

- Might break terms of service (robots.txt)
- Not as “clean” as an API
- Very fragile! We are depending on implementation details for our solution to work.

1 - Programming Library Classes

General Idea

Write functions/classes that implement a specification like some library function/class that we have used before.

You don't have to memorize every function/class we have used, we will describe the behavior

Resources

- Lecture 10
- Lectures 11-12 can be helpful, but a bit too focused on NLP
- Section 4
- HW4 (focus on classes and objects, less TF-IDF and search)

2- Miscellaneous

General Idea

A series of short-answer questions about topics we have discussed in the course. Topics include but are not limited to

- Ethics
- Hashing
- Machine Learning
- Testing

Resources

- Lecture 16, 21, 24
- Homeworks

3 - Geospatial + Tabular Data

General Idea

Write code that interacts with pandas and geopandas objects

Resources

- Lectures 19+20
- Section 7
- HW6

4 - Image Processing

General Idea

- Simulate operations like broadcasting or convolutions
- Write code to do some image manipulation
- Write code that does something like a convolution

Resources

- Lectures 23+23
- Section 8
- HW7

Poll Everywhere

Think

1.5 minutes

pollev.com/cse163

Problem 1: How many weights will the neural network have to learn?

Poll Everywhere

Pair

2 minutes

pollev.com/cse163

Problem 1: How many weights will the neural network have to learn?